P.v.d.Lippe: Manuskript meiner Buchbesprechung von

Thomas P. Ryan, Sample Size Determination and Power, John Wiley, Hoboken, N.J. 2013

erschienen in Jahrbücher für Nationalökonomie und Statistik, Bd. 235, Heft 2 (Januar 2015)

Jeder Statistiker kennt das, dass die Frage "Wie groß muss meine Stichprobe sein, damit sie repräsentativ ist?" wohl die Frage ist, die einem als Statistiker am häufigsten gestellt wird. Hier haben wir jetzt ein ganzes Buch, das allein diesem Thema, also dem Stichprobenumfang gewidmet ist ("Repräsentativität" ist an keiner Stelle des Buches ein Thema). Das ist deshalb besonders verdienstvoll, weil sich die meisten Anwender bestenfalls an ein relativ kurzes Kapitel über den "mindestens notwendigen Stichprobenumfang n" in der entsprechenden Statistikvorlesung erinnern und weil sie deshalb wenig in der Hand haben, um die Frage zu beantworten. Meist beschränkt man sich dabei auf die Formel

$$(1) \qquad n \geq \left(\frac{z_\alpha \sigma}{e} \right)^2 ,$$

die aber nach Ryan oft zur Unterschätzung von n führt und bei der mit e, der halben Breite des symmetrischen (1 - α)-Konfidenzintervalls für μ (dem "desired margin of error", oder "maximum error of estimation") und mit z_α, dem zu 1- α gehörigen Perzentil der Standardnormalverteilung (etwa $z_\alpha = 1{,}6445$ bei 1- $\alpha = 0{,}95$) zwei Qualitätsanforderungen an die Stichprobe gestellt werden. Sehr viel mehr noch als diese beiden "Anforderungen" ist aber die eigentlich problematische Größe (der nuissance parameter) σ^2, die Varianz von x in der unbekannten Grundgesamtheit, weil sie *Annahmen* erforderlich macht, die realistisch sein sollten.

Weniger üblich als (1) ist es, die Formel

$$(2) \qquad n \geq \left(\frac{(z_\alpha + z_\beta)\sigma}{(\mu_0 - \mu_1)} \right)^2 ,$$

zu erwähnen, die sich nicht an der gewünschten Qualität eines Konfidenzintervalls orientiert, sondern an der von Tests einer Hypothese über μ. Hier sind bei der Bestimmung von n noch mehr Entscheidungen erforderlich, und zwar nicht nur über σ, sondern auch über die für erforderlich gehaltene Power 1 - β und die hierfür bestimmende Differenz $\mu_0 - \mu_1$ (zwischen H_0: $\mu = \mu_0$ und der spezifizierten Alternativhypothese H_1: $\mu = \mu_1$). Beide Vorgaben können in der Praxis stets nur mehr oder weniger, i. d. R. aber wohl eher weniger fundiert sein. So wie es meist nicht mehr als nur eine Konvention ist, für α den Wert 5% anzusetzen (Ryan zitiert einen seiner Lehrer, der dies damit begründete, dass man eben auch fünf Finger hat) so ist es wohl auch nur eine Konvention, die gewünschte Power mit 80% (also 1 - $\beta = 0{,}8$) anzusetzen.

Beide hier zitierte Formeln betreffen die Inferenz über Mittelwerte (μ) und Anteilswerte (π). Der Stichprobenumfang bei solchen Fragestellungen sowie bei der Inferenz über Varianzen, zwei und mehr Anteilswerte (proportions) einschließlich Vierfelderkorrelationen ist Gegenstand von Kapitel 2 bis 4, was zusammen mit dem sehr verdienstvollen Kapitel 1 "Brief Review of Hypothesis Testing Concepts/Issues and Confidence Intervals" fast 40% (über 140 Seiten) des Buches ausmacht. Diese Gewichtung scheint mir sehr im Einklang mit der Relevanz für die Praxis von Stichprobenuntersuchungen zu sein.

Die übrigen acht Kapitel betreffen neben Regression und Korrelation, Experimental Designs und Qualitätskontrolle (control charts) auch sehr spezielle Themen wie klinische (Kohorten-) Versuche (Zeitschriften aus diesem Bereich, also aus der medizinischen und biologischen Statistik dominieren auch die mehr als reichlich gegebenen Literaturhinweise) sowie nichtpa-

rametrische und multivariate Methoden. In diesen Kapiteln werden oft die relevanten Formeln nicht mehr so ausführlich hergeleitet und kommentiert, wie z.B. (1) und (2).

Wie erwähnt, halte ich das erste Kapitel, auch wenn es relativ kurz ist, angesichts der vielen Missverständnisse über p-values, Power usw. unter den Statistikanwendern für sehr verdienstvoll. Es wird u.a. überzeugend dargelegt, warum bei n sowohl ein "zu groß" als auch ein "zu klein" ein Problem ist, dass n bei anspruchsvolleren Vorgaben bezüglich α und 1 - β rapide in die Höhe schnellt, dass H_0 fast immer falsch ist, so dass ein Verwerfen von H_0 als solches nicht sehr sensationell ist und vor allem warum – unter Berufung auf Hoenig und Heisey – die sog. "observed" oder "retrospective power" (man legt in Ermangelung von inhaltlich fundierten Gründen für die Wahl eines konkreten Zahlenwerts für μ_1 einfach den gefundenen Stichprobenwert, also $\bar{x}$ der Berechnung von 1 - β zugrunde) unsinnig ist; denn sie liefert keine über den durch $\bar{x}$ bestimmten p-Wert hinausgehende Information. Die gut verständliche Darstellung wird (nicht nur in diesem Kapitel, sondern generell) ergänzt durch (Rechen-)Beispiele und "exercises", darunter auch einfach erscheinende verbale Fragen, wie etwa "Why would you not want to have a hypothesis test that has high power (such as .90) for detecting a very small difference in two population means?", oder warum die kurze Definition "Power is the probability of observing the smallest effect of clinical interest, if it exists in the population" noch einer Ergänzung bedarf (S. 14f.).

Auch an späterer Stelle wird immer wieder betont, dass bei der Bestimmung von n *neben*, wenn nicht gar *vor* den oft wenig fundierten Erwägungen über die gewünschte Power und die absolute bzw. relative, d.h. durch σ dividierte Effektstärke $\mu_0 - \mu_1$, also was als ("clinically meaningful" (S. 147) nachweisbar sein soll[1] ganz andere Kriterien zu bedenken sind, nämlich Kosten, Zeitaufwand und ethische Aspekte. Das Wort "Power" erscheint mit Recht im Titel des Buches; denn es wird immer wieder problematisiert, z.B. in Kap. 7 im Zusammenhang mit klinischen Versuchen wegen der gerade dort sehr offensichtlichen ethischen Aspekte. Es ist bemerkenswert, dass die "contention that inadequate power makes a study unethical" sowohl zur Ablehnung von "under-powered" Studien auf Basis der "threshold myth, a false belief that studies with less than 80% power cannot be expected to produce enough scientific or practical value to justify the burden imposed on participants"(S. 252) als auch die gegenteilige Position, wonach (auch aus ethischen Gründen) eine zu große Power und ein zu großes n abzulehnen ist. Nach Ryan ist zwischen assumed (asp), actual (acp) und target power (tp) zu unterscheiden. Die asp wäre gleich der acp wenn die *angenommenen* Werte der nuissance parameter (wie σ in (*) und (**) oder σ_1 und σ_2 im Zwei-Stichproben-Fall) gleich den *tatsächlichen* Werten wären (S.59f). Die asp bzw. tp ist fast nie gleich der acp und "power will always be unknown both before and after a study has been conducted" (S. 253).

Schon bei der einfachen Formel (*) wird ein Grundproblem deutlich: die in die Formel einzusetzende Größe σ ist nicht eine unproblematische *Annahme*, sondern genau genommen ein (wie μ) zu schätzender Parameter, also (als Schätzwert $\hat{\sigma}$) eine *Zufallsvariable*. Das gilt auch für die mit σ zusammenhängenden Power. Also müssten eigentlich Konfidenzintervalle für σ, bzw. bei der multiplen Regression mit K Regressoren (in der Praxis von Nicht-Experimentdaten meist Zufallsvariablen) für die Varianzen von $x_1, \ldots, x_K$ und die Power bestimmt werden, und das alles "would be very complex and perhaps even intractable when confidence bounds (or a joint confidence region) must be developed for handling multiple unknown parameters … Consequently, what some discerning readers might call a naïve approach of substituting values for unknown parameters will continue to be used throughout this book because that is simply all that is available "(S. 70). Man sieht also, und darauf weist Ryan immer

[1] Es geht also darum, was als *inhaltlich*, *substanziell* "signifikant" angesehen wird, aber "…there is no a priori reason why one specific value of a difference … is worthy of detection"(S. 39). Es ist also schon schwierig die zur richtigen Bemessung der Power erforderliche Effektgröße zu definieren.

wieder hin, dass die Bestimmung von n selbst bei ziemlich komplizierten Ansätzen immer noch als eine notwendig etwas vergröberte Betrachtung angesehen werden muss: "Sample size determination is never going to be an exact science, so being overly concerned with 'exactness' in certain ways seems inappropriate" (S. 115). Hinzu kommt, dass nicht auszuschließen ist, dass "Peer reviewers often make unfounded criticisms" was n und die Power betrifft (S. 243) und dass wir in Nachhinein "could not directly assess whether assumptions had been manipulated to obtain feasible sample sizes" (S. 259).

Dass hier die Dinge sehr schnell kompliziert werden mag auch der Grund dafür sein, dass die Bestimmung des Stichprobenumfangs n schon auch gelegentlich als ein gemessen am Nutzen zu schwieriges und zu aufwändiges Unterfangen in Misskredit geraten ist (S. 37)[2] und in der Praxis meist einem der zahlreichen Computerprogrammen überlassen wird (über deren Möglichkeiten und Grenzen in diesem Buch sehr viel die Rede ist). Es ist zu befürchten, dass diese aber quasi als black box benutzt werden, was dann auch Zweifel aufkommen lässt ob die Relevanz von empirischen Befunden richtigen eingeschätzt wird. Es ist deshalb sehr zu loben, dass Ryan nicht nur wiederholt verlangt, dass Forscher erklären, wie sie zu ihrem Wert von n gelangt sind (z.B. auf S. 57) sondern auch immer wieder auf generelle Beschränkungen hinweist, wie z.B. schon auf S. 47 "As when any statistical tool is used, the results will be only as good as the assumptions that are made. These types of problems do not render sample size determination useless, but experimenters should keep in mind that the specified power used in sample size determination is not going to be the actual power that a study has. (There is always uncertainty in statistics because random variables are involved.)"

Weil Testen mehr als Intervallschätzen im Fokus der meisten Betrachtungen zur Bestimmung von n steht, war es auch geboten, Alternativen zum Neyman-Pearson ("frequentist") Ansatz kurz darzustellen. In mehreren Kapiteln werden – jeweils etwas stiefmütterlich und mit Skepsis – "Bayesian Approaches" erwähnt. Andere Ansätze kommen noch kürzer zur Sprache, z.B. wird der angesichts der unten zitierten Arbeit von Strug et al. nicht uninteressante "evidential approach" nur mit gerade mal vier Zeilen bedacht (S. 37).

Kommt man von den Kapiteln über Mittel- und Anteilswerte zu den späteren (weiterführenden) Kapiteln, so wird die Bestimmung von n oft sehr schwierig. Der Grund ist nicht nur, dass die entsprechenden Formeln für n oft viel mehr Vorgaben erfordern als Formeln wie (1) und (2), sondern auch, dass vorzugebende Parameter (wie etwa die β-Koeffizienten bei der Regression) oft keine evidente *inhaltliche* Bedeutung haben und dass deshalb auch konkrete *zahlenmäßige* Vorgaben (z.B. für the smallest acceptable R^2) schwer zu begründen sind, und diese auch untereinander zusammenhängen können (etwa R^2 wegen $R^2 = f^2/(1+ f^2)$ mit der effect size f und damit auch mit der Power). Hinzu kommt z.B. Multikollinearität und etwas, was bei Schätzen und Testen im Falle von μ und π kein Problem ist, dass nämlich auch die mit der Stichprobe verbundene Aufgabenstellung ein Thema sein kann (ist x_k in $\hat{y}_i = \alpha + \beta k_{ki}$ der "richtige" Regressor?). So gibt es hier Faustregeln, die für Ryan durchaus akzeptabel sind angesichts der Komplexität der Bestimmung von n (S. 150), wie z.B. die Regel "10 Beobachtungen pro Regressor" (nach Draper and Smith) oder "the model is useful only if the value of the t-statistic…is at least twice the critical value" (nach Wetz S. 147), was immer auch "useful" heißen mag. Relativ viel Platz wird der logistischen Regression gewidmet, bei der die Bestimmung von n noch komplizierter ist als sonst in der Regressionsanalyse (S. 156, 159). Es ist deshalb nützlich, dass Ryan oft angibt, welche Software welche Formel benutzt, auch wenn die Formeln selbst oft quasi vom Himmel fallen.

[2] Es könnte auch der Grund dafür sein, dass – wie erwähnt – trotz großer praktischer Bedeutung diesem Thema in Einführungsvorlesungen meist nicht sehr viel Platz eingeräumt wird.

Auf die "späteren" Kapitel (ab Kap. 6 "Experimental Designs"), kann hier aus Platzgründen nicht näher eingegangen werden. Das Buch ist sehr umfassend, es bietet einen sehr gelungenen, nützlichen, aber auch nicht wenig anspruchsvollen Gesamtüberblick und es enthält, wie gesagt, auch eine enorme Fülle von (gut über 900) Literaturhinweisen. Es will sowohl Nachschlagewerk als auch Lehrbuch sein, was dem Autor auch sehr gelungen ist. Benutzt man es als letzteres, ist die Lektüre jedoch manchmal etwas ermüdend, weil oft – ebenfalls aus Platzgründen – statt Details ausführlich genug in eigenen Worten zu präsentieren, nur lange Auflistungen von Namen und Jahreszahlen geboten werden.

Hoenig, J. M., and D. M. **Heisey**, The abuse of power: The pervasive fallacy of power calculations for data analysis, The American Statistician, 55/1 (2001), 19 – 24.

Strug, L. J., C. A. **Rohde**, and P. N. **Corey**, An introduction to evidential sample size calculations, The American Statistician, 61/3 (2007), 207 – 212.